Olympus: an Ambient Intelligence Architecture on the Verge of Reality*

F. Bertamini R. Brunelli O. Lanz A. Roat A. Santuari F. Tobia Q. Xu
ITC-irst
Via Sommarive 18
38050 Povo, Italy
brunelli@itc.it

Abstract

This paper presents Olympus, a modular processing architecture for a distributed ambient intelligence. The system is aimed at detailed reporting of people wandering and gesturing in complex indoor environments. The design of the architecture has been driven by two main principles: reliable algorithm testing and system scalability. The first goal has been achieved through the development of Zeus, a real time 3D rendering engine that provides simulated sensory inputs supported by automatically generated ground truth for performance evaluation. The rendering engine is supported by Cronos, a flexible tool for the synthesis of choreographed motion of people visiting museums, based on modified force fields. Scalability has been achieved by developing Hermes, a modular architecture for multi-platform video grabbing, MPEG4 compression, stream delivery and processing using a LAN as a distributed processing environment. A set of processing modules has been developed to increase the realism of generated synthetic images which have been used to develop and evaluate algorithms for people detection.

1. Introduction

Vision-based perception of moving people, the analysis of the resulting motion patterns and the understanding of their gestures has been and still is one of the most active research fields in Computer Vision [1, 2]. This continuous interest is sustained on one side by the number and importance of potential applications, on the other side by the strong challenges posed by the development of algorithms for robust tracking of people in spite of occlusions and dynamic lighting. While significant efforts have been spent in the development of algorithms and architectures, extensive performance evaluation of the resulting systems while a key methodological issue has not been given proper attention. This paper presents the structure of Olympus, an ambient intelligence architecture relying on multiple computer controllable cameras to track moving, partially (self) occluded articulated objects and determine the position of selected skeletal points in space-time. Olympus has a scientific target and a methodological one:

- the development of efficient algorithms for the adaptive management of a population of cameras monitoring complex multi-room environments, cooperating in people tracking and occlusion disambiguation;
- the introduction of rigorous assessment of algorithm performance through the generation of synthetic video sequences of graded difficulty associated with computer generated ground truth for evaluation (a *perceptual oracle*).

The currently envisaged application of Olympus is the development of smart museum environments but the resulting technological framework will provide the building blocks for several real world applications such as: smart rooms and environments (museums, video conferencing rooms), advanced video surveillance systems, monitoring of physical rehabilitation sessions, and automatic scoring of divings and similar activities.

The architecture of Olympus is presented in Section 2 and its *perceptual oracle* is described in Section 3. Synthesized choreographed people motion is presented in Section 4. Distributed sensor input and processing is addressed in Section 5, while first results on the use of algorithm evaluation using Zeus are reported in Section 6.

2. Olympus

Monitoring people movements in complex environments, analyzing the resulting motion patterns and understanding people gestures corresponds to a high level of visual competence that can most appropriately be identified as

*Research partly funded by Provincia Autonoma di Trento under Project PEACH: Personalized Experience of Active Cultural Heritage.

ambient intelligence. It is then natural to define the Olympus architectural details in terms of the input sensors, the active (processing) transmission channels on which the information flows, and the *brain* interpreting the view of the external world presented by the sensors. The overall architecture is presented in Figure 1 where we can identify the following modules:

Zeus/Hera : the graphical reality engines providing annotated, artificial, sensory inputs to the system;

Gaia : the actual sensor system, based on real cameras and depth sensors;

Cronos : the people motion simulator, providing realistic group dynamics;

Hermes : the network transporting the signals from the sensors to the interpreting brain Athena;

Athena : the *ambient intelligence* interpreting multiple data flows from the real and simulated environments.

A characterizing feature of the proposed architecture is the presence of two different sensor systems: Gaia and the Hera/Zeus couple. The latter is first used during the nurturing phase of algorithm development. The graphical simulators ease the investigation of complex architectures by providing simulated sensory input supported by complete knowledge of the corresponding environment. They can be used to assess algorithm performance by acting as *perceptual oracles* to which computed data must be compared. The availability of synthetic sensory input also supports extended experimentation with environments of graded complexity, different sensor layouts and even the simulation of new sensors, such as laser telemeters with given resolution. Gaia will be the main sensor systems in the real environment but the Hera/Zeus duo will still have its function. Active sampling of the environment by the Olympus sensor network will enable the reconstruction of its geometry and textural properties. This information can be used by the graphical simulators to synthesize specific views on demand, implementing an active memory useful for the interpretation of environment dynamics.

The following sections describe with some detail the different modules with the exception of Gaia: the physical sensor network will be designed and realized after extensive simulation based on rendered images.

3. Zeus and Hera

The usage of simulated sensory inputs is a major feature of Olympus architecture and is realized by two rendering engines, Zeus and Hera, respectively aimed at fast and

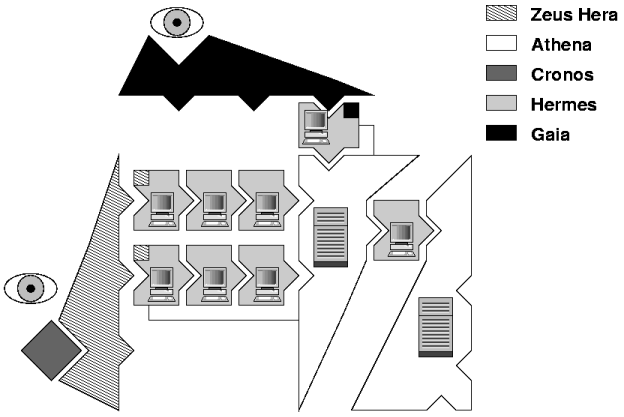

Figure 1. The architecture of the Olympus visual system: a set of modules provide real (Gaia) or simulated (Zeus/Hera) sensor data which are then transmitted (and possibly partially processed) over a network (Hermes) to the interpreting ambient intelligence (Athena), itself distributed over multiple processing units.

physically correct rendering. They transform a numeric description of the geometry and materials of a dynamic environment into realistic digital images that can be used for the development and test of tracking algorithms (see Figure 2). Zeus is meant to be a fast renderer, taking advantage of state of the art graphic boards and 3D game technologies [9]. He is then based on the widely adopted OpenGL standard and currently features:

- precomputed global illumination solution for the static environment and dynamic lighting of moving objects;
- computation of moving objects shadows for static colored lights, with optional soft shadowing effects;
- efficient rendering of complex indoor environments using Binary Space Partition trees and Potentially Visible Sets;
- articulated character rendering with the possibility of switching among different motion classes (such as *walking* and *running*) with automatic interpolation;
- a client/server architecture, whereby distributed clients can request a view of the simulated world from any position, supported by the corresponding *perceptual oracle* information.

The alternative rendering engine, Hera, aims at physically correct image generation. She is currently based on the

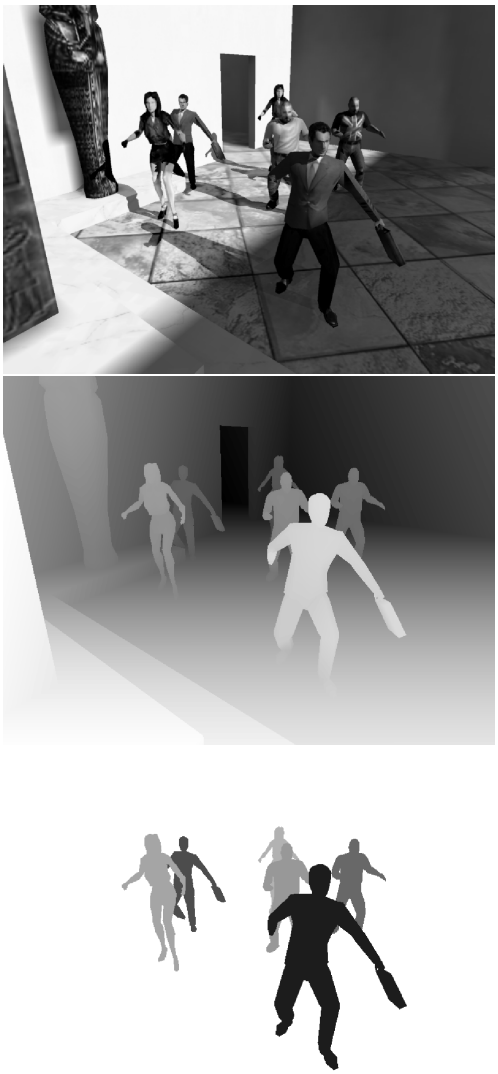

Figure 2. The information generated by Zeus: the visual stimuli, a depth map, and the 'oracle' identifying each person with a different label.

open source lighting simulation and rendering system Radiance [4], which has been used in the past for several architectural simulators. Radiance is able to support complex geometries and materials, and both direct and indirect illumination providing physically accurate images. Currently, a rendering API has been developed that provides on a pixel per pixel basis information on the actual 3D world coordinates, object and material identifiers, surface normal and radiance values. An important aspect of Hera API is the support for fine tuning of the rendering quality parameters with a pixel granularity. This feature permits foveated rendering of image regions of particular interest for Athena, the ambient intelligence sensing the environment.

4. Cronos

The generation of appropriate motion patterns for the simulated characters moving in the virtual environment is the next issue to be solved in order to provide useful synthetic inputs to the system. Although rendering the characters is by itself a complex task, the creation of a suitable choreography of groups moving convincingly in the environment poses additional challenges. Cronos, a special client of the architecture graphical simulator, is able to animate groups of simulated characters. It is based on a hybrid approach relying on flock simulation of internal group dynamics and on a novel version of computational fields [5]. The reason for the hybrid approach to simulated people motion synthesis stems from the fact that, while plausible flock behavior naturally emerges from the simple rules proposed in [7], consistent goal oriented behavior can not be easily created in the same way. The solution employed by Cronos is indirect groups control through invisible leaders: while each element of a group is attracted by the position of its leader with a force proportional to its distance, the motion of the leader itself characterizes the goal oriented motion of the group. The leaders can then be moved on a graph whose nodes represent physical locations in the museum. The nodes corresponding to the exhibits are given a set of descriptive labels from which an interest score is derived based on the group profile while the arcs are labeled with the corresponding spatial distances. In this case, people motion can be planned in advance, optimizing the fruition of the exhibits (sum of the scores) subject to predefined temporal constraints on the length of the visit. However, such an approach is not completely satisfactory as people do not, usually, optimize their visit this way. Furthermore, the impact of local events (such as overcrowding) on group trajectories can not be modeled with the necessary spatial detail unless a large number of nodes is inserted into the graph. The solution proposed by Cronos is based on a novel version of the abstract computation field concept introduced by Mamei and colleagues [5] for the coordination of distributed agents. In the original formulation, each agent perceives the position of the other agents through a force field and a coordinated behavior emerges from the agents following the force field and dynamically re-shaping it by changing their position. A major limitation of this approach in the context addressed by Cronos derives from the effect of complex environment geometry. Ignoring environment geometry in the generation of the fields originates strange motion patterns as characters motion cannot ignore existing obstacles. Modeling walls and similar constraints with additional fields may

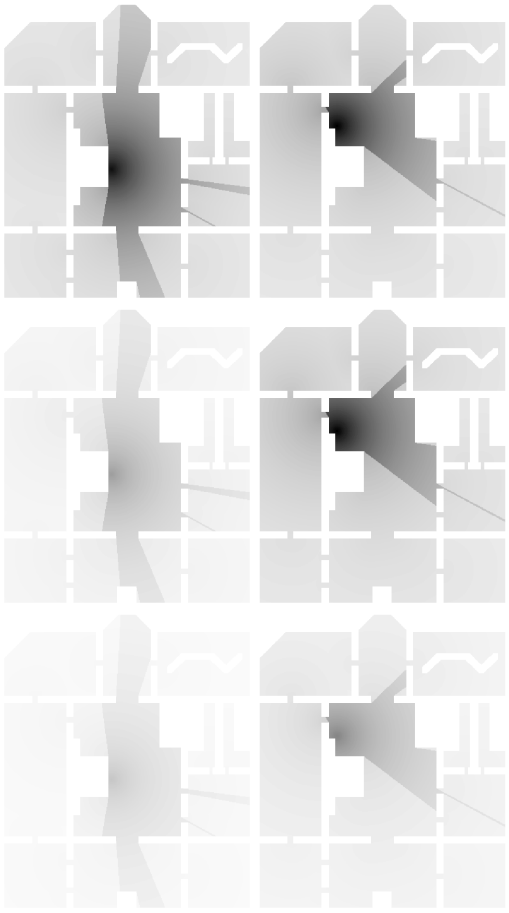

Figure 3. The two columns reports the evolution in time of the co-fields originated from two different exhibits: the darker the field the stronger the attraction of the exhibit. The visitor is at first attracted by the one corresponding to the left. As time pass by, she gets less and less interested in it till her attention is captured by the nearby exhibit. Field discontinuities are due to the depression of secondary source emission, capturing the effect out-of-sight out-of-interest.

also result in strange motion patterns that prevent the agents from reaching the expected objectives. The version of co-fields used by Cronos is generated by propagating *interest rays* from each exhibit. The rays move linearly in a discretized space and mark each floor cell with the minimum distance to the exhibit and the coordinates of the current ray source. In order to cover the whole environment (in a way not dissimilar to diffuse illumination) each pixel acts as a new source. The resulting fields incorporate geometrical information and can be effectively modulated by the complexity of the path leading to the exhibit as well as the mere distance to be traveled (see Figure 3). The implemented system permits the simulation of groups with different interests in museum exhibits as well as different visiting strategies (detailed planning, casual, time limited etc.) by appropriate modulation of co-fields.

5. Hermes

The necessity of integrating multiple information flows from a network of sensors and to perform a substantial amount of computation suggested from the very beginning the development of Hermes, a distributed architecture for image delivery and processing. He currently provides a multi platform environment for the acquisition of visual information, its compression according to the MPEG4 standard, and its transmission and processing using multiple distributed modules connected in a pipeline fashion via TCP/IP. Specialized modules can be used at the sensor interfaces, to control real (Gaia) and virtual (Zeus/Hera) cameras, and to insert the corresponding data in the processing network leading to the Athena core. This architecture enables efficient transmission of video data over low bandwidth channels, permitting the use of a wireless infrastructure even for multiple input channels. The current implementation is used to increase the realism of the images produced by Zeus by adding different kinds of noise and optical effects (see Figure 4) mimicking real cameras characteristics:

Depth of field mimicking the loss of detail for objects far away from camera focus: it is simulated using depth information provided by Zeus with every frame.

Motion blur, due to slow shutter time with respect to scene motion: it is realized by shutter time integration of a temporally super-sampled sequence generated by the graphical simulator.

Interlace, due to some devices reconstructing the image by drawing the even lines before the odd ones. The generated frame can then be affected by asynchronism, in which half of the image belongs to a previous frame.

Impulse and Gaussian noise, which respectively degrade the image by saturating the value of a given percentage of pixels and induce a granular appearance.

Current ongoing development aims at refining the module control architecture. Easy deployment of the processing modules will be supported by graphical tools able to *compile* an iconic definition of the processing pipeline onto a set of workstations locally connected via gigabit Ethernet.

Figure 4. Hermes can be used as a flexible distributed processing pipeline to apply video-post effects to the synthetic images produced by Zeus increasing their realism: (from left to right, top to bottom) motion blur, depth of field simulation, interlacing and impulse noise.

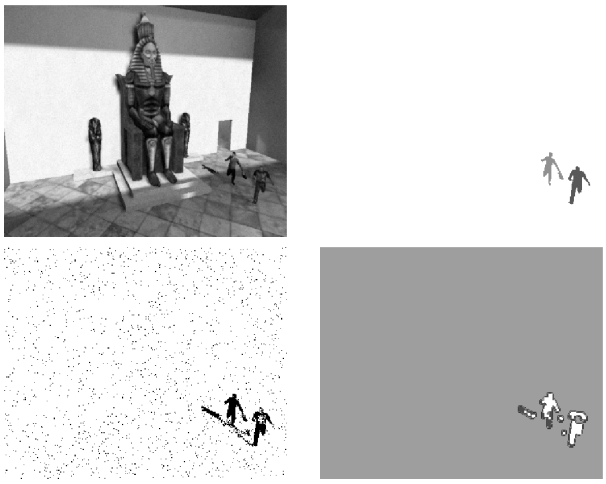

Figure 5. Some preliminary examples of Athena at work. From top to bottom, left to right, we have an image provided by Zeus and contaminated with Gaussian noise added by the Hermes processing network, the corresponding Zeus 'oracle' response, providing pixel level ground truth, the first step in segmenting moving people from the environment, and the final version showing the actual segmentation and the estimated shadows.

6. Athena

The simulated and real sensorial input streams of Olympus must be integrated and interpreted providing a high level description of the environment. Raw sensor information is progressively processed along the transmission paths, extracting features useful for higher level interpretation. This final step is carried out by Athena, the ambient intelligence linked to the multi sensor monitoring network. The development of Athena monitoring competence is currently based only on simulated input provided by Zeus. Upon reaching acceptable competence on a suite of synthetic tests of increasing difficulty, Athena will face the real world where the population of cameras simulated by Zeus will be replaced by a network of real consumer cameras, Gaia. Athena's visual sensors are actively controlled to provide the maximum possible information on the monitored environment through reactive connection of multiple sensors into surveillance agencies. These agencies cooperate in resolving interpretation ambiguities due to occlusions and poor camera resolution.

Athena is currently able to discriminate people from a static background and to identify people shadows, providing accurate silhouette information. This functionality is based on a simplified two steps tracking algorithm. The first *learning* step consists in the estimation of a reference background image, computed over a number of static background frames (e.g. a view of the museum interiors without visitors). Each image pixel $(I_R(i), I_G(i), I_B(i))$ is represented by its expected mean color value $(\mu_R(i), \mu_G(i), \mu_B(i))$ and standard deviation $(\sigma_R(i), \sigma_G(i), \sigma_B(i))$, hue, saturation, and color brightness distortion [3]:

$$\alpha_i = \frac{\left(\frac{I_R(i)\mu_R(i)}{\sigma_R^2(i)} + \frac{I_G(i)\mu_G(i)}{\sigma_G^2(i)} + \frac{I_B(i)\mu_B(i)}{\sigma_B^2(i)}\right)}{\left(\left[\frac{\mu_R(i)}{\sigma_R(i)}\right]^2 + \left[\frac{\mu_G(i)}{\sigma_G(i)}\right]^2 + \left[\frac{\mu_B(i)}{\sigma_B(i)}\right]^2\right)}$$

representing the projection onto the *chromaticity line* of the current pixel. The second *classification* step consists of a preprocessing stage, binarization, and pixel classification. A first rough discrimination objects/background is made through *background subtraction* relying on pixel statistics computed in the first step. Each pixel is then classified as *foreground* or *background* depending on whether its difference from the reference background image is statistically significant or not. The resulting (noisy) image is cleaned by morphological erosion filtering. At the same time, brightness distortion is computed and color space conversion, from RGB to HSV, performed. Finally, each foreground pixel is given a score based on the result of a set of three

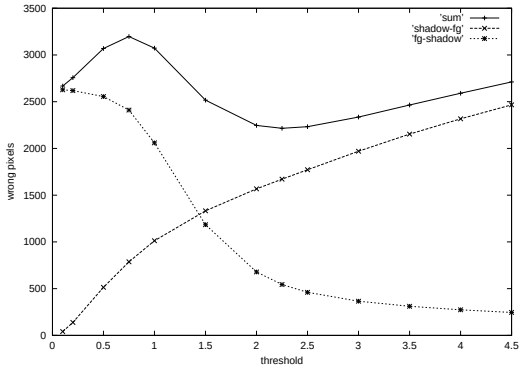

Figure 6. The plot reports the error made by the current Athena algorithm for people/background segmentation. After segmenting the image into background and foreground plus shadows, the algorithm tries to identify shadows using a set of classifiers combined through scoring. Foreground/shadow discrimination of component classifier C1 is based on a single threshold and the plot reports the error as a function of its value. Manual computation of algorithm performance with this detail and accuracy would be extremely costly.

classifiers:

C1 computes the ratio between the saturation of the current image and the corresponding value of the background image: whenever the result is greater than a parametric threshold, one point is assigned;

C2 compares the brightness distortion of the current image with two thresholds, one point is given when the result falls within;

C3 the hue difference between the current image and the reference one gives one point if below threshold.

A pixel is labeled as shadow or foreground, if the resultant score is respectively greater or lower than a predefined threshold (see also Figure 5, alternative methods for shadow detection can be found in [6]). The perceptual oracle functionality provided by Zeus enables very accurate performance assessment of the algorithm and some results are reported in Figure 6.

7. Conclusions and future work

We have introduced Olympus, a flexible architecture based on the use of graphical simulators with perceptual oracle functionalities, for the evaluation of *smart environment* systems. Olympus has several novel features such as Zeus, a high quality renderer providing synthetic images and perceptual oracle functionalities, Cronos, a flexible puppeteer for the synthesis of choreographed motion of people visiting museums, Hermes, a distributed and multi platform architecture for the acquisition, compression and processing of video streams, and Athena, a currently evolving set of algorithms for ambient intelligence that will be thoroughly evaluated in environments of graded complexity using Zeus generated ground truth.

Current activity is mainly focused on improving the visual competence of the interpreting module, Athena, with the introduction of robust algorithms for people tracking and gesture recognition based on cooperative multi camera operation. Automatic reconstruction of the monitored environment will be used to provide increasingly realistic synthetic environments to be generated by the graphical simulators Zeus and Hera.

Olympus will also provide the basis for further pursuing the approach pioneered by Terzopoulos [8] to investigate evolutionary development of visual competence in a distributed processing environment and, more generally, the possibilities offered by supervised learning in the field of visual-like sensory signal processing.

References

[1] S. L. Dockstader and A. M. Tekalp. Multiple Camera Tracking of Interacting and Occluded Human Motion. *Proc. of the IEEE*, 89(10), 2001.

[2] I. Haritaoglu, D. Harwood, and L. S. Davis. W^4: Real-Time Surveillance of People and Their Activities. *IEEE Trans. On PAMI*, 22(8):809–830, 2000.

[3] T. Horprasert, D. Harwood, and L. S. Davis. A Robust Background Subtraction and Shadow Detection. In *Proc. ACCV'2000*, Taipie, Taiwan, January 2000.

[4] G. W. Larson and R. A. Shakespeare. *Rendering with Radiance The Art and Science of Lighting Visualization*. Morgan Kaufmann Publishers, 1998.

[5] M. Mamei, F. Zambonelli, and L. Leonardi. Co-Fields: A Unifying Approach to Swarm Intelligence. In *Proc. of 3rd International Workshop on Engineering Societies in the Agents' World*, Madrid, Sept. 2002.

[6] A. Prati, I. Mikic, C. Grana, and M. M. Trivedi. Shadow Detection Algorithms for Traffi c Flow Analysis: a Comparative Study. In *Proceedings of IEEE Intl. Conference on Intelligent Transportation Systems*, pages 340–345, 2001.

[7] C. W. Reynolds. Flocks, Herds, and Schools: A Distributed Behavioral Model. *Computer Graphics*, 21(4):25–43, 1987.

[8] D. Terzopoulos and T. F. Rabie. Animat vision: Active vision in artifi cial animals. *Videre*, 1(1), 1997.

[9] A. Watt and F. Policarpo. *3D Games Real-time Rendering and Software Technology*, volume 1. Addison-Wesley, 2001.